Coming Soon: the new CIEMPIESS datasets for speech recognition in Mexican Spanish

Carlos Daniel Hernández Mena

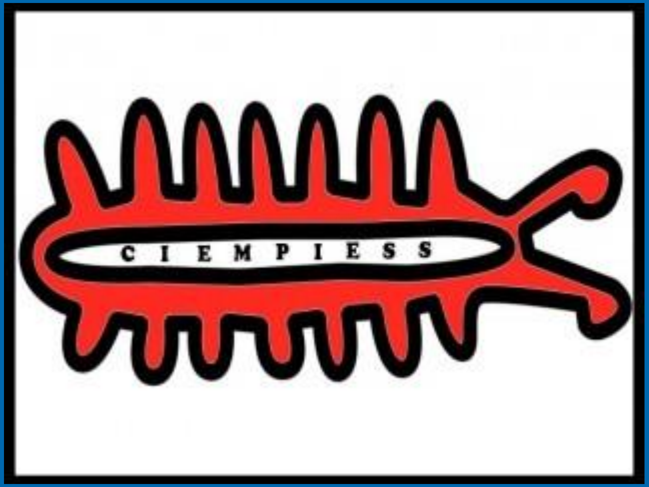
The CIEMPIESS Corpus

(The Original CIEMPIESS)

Characteristics

- Published at LDC in June 2015 (LDC2015S07).
- 17 hours of train and 1 hour of test
- Word level timestamps (PRAAT)
- Includes files to perform experiments with the CMU-SPHINX3
- Creative commons share-a-like license

Did you know?

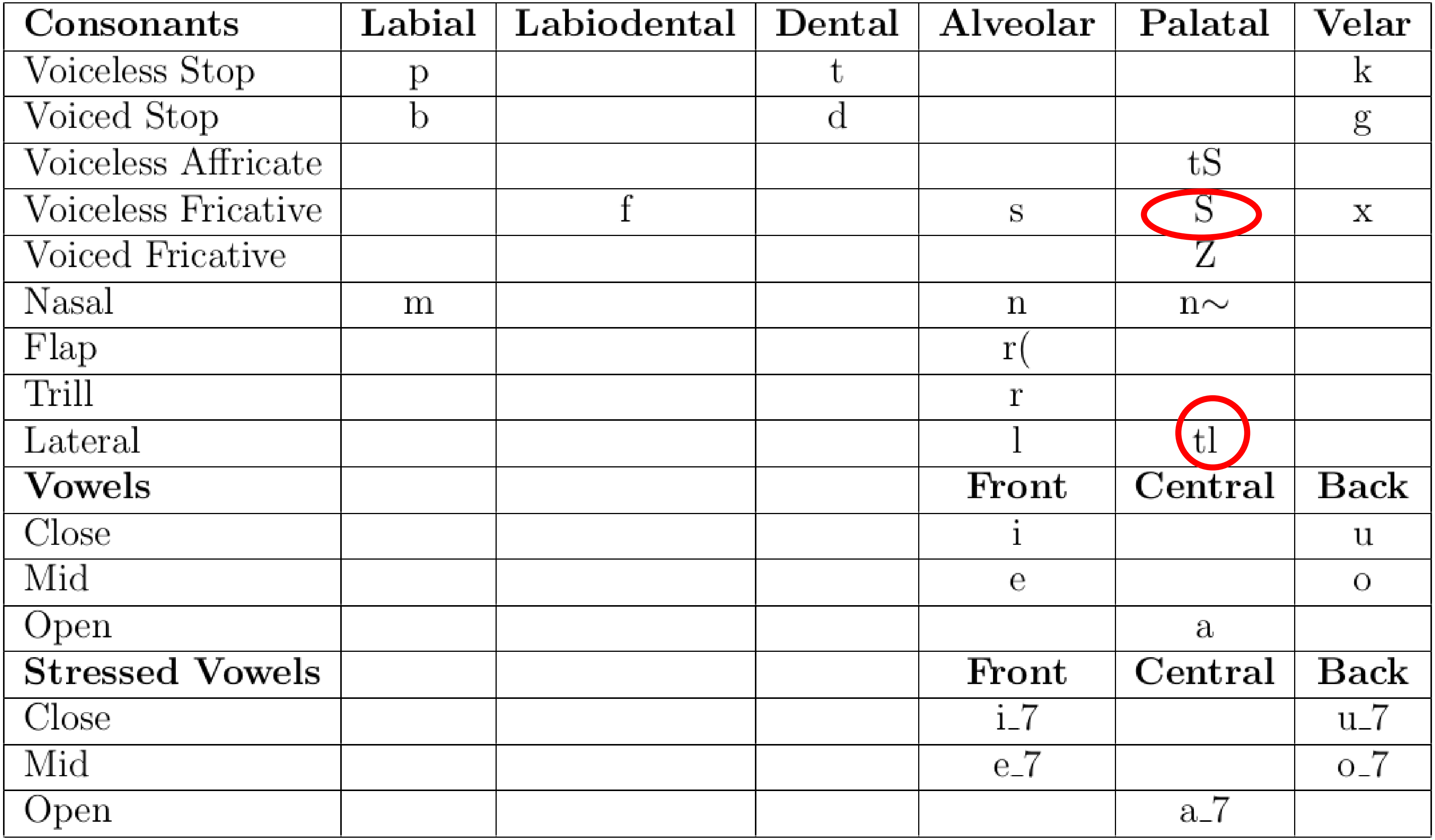


In Spanish, “CIEMPIESS” sounds like “cienpiés” (“centipede” in English)

“CIEMPIESS” is the acronym for: Corpus de Investigación en Español de México del Posgrado de Ingeniería Eléctrica y Servicio Social

Image from “Flora y Fauna, mayo 27, 2017”

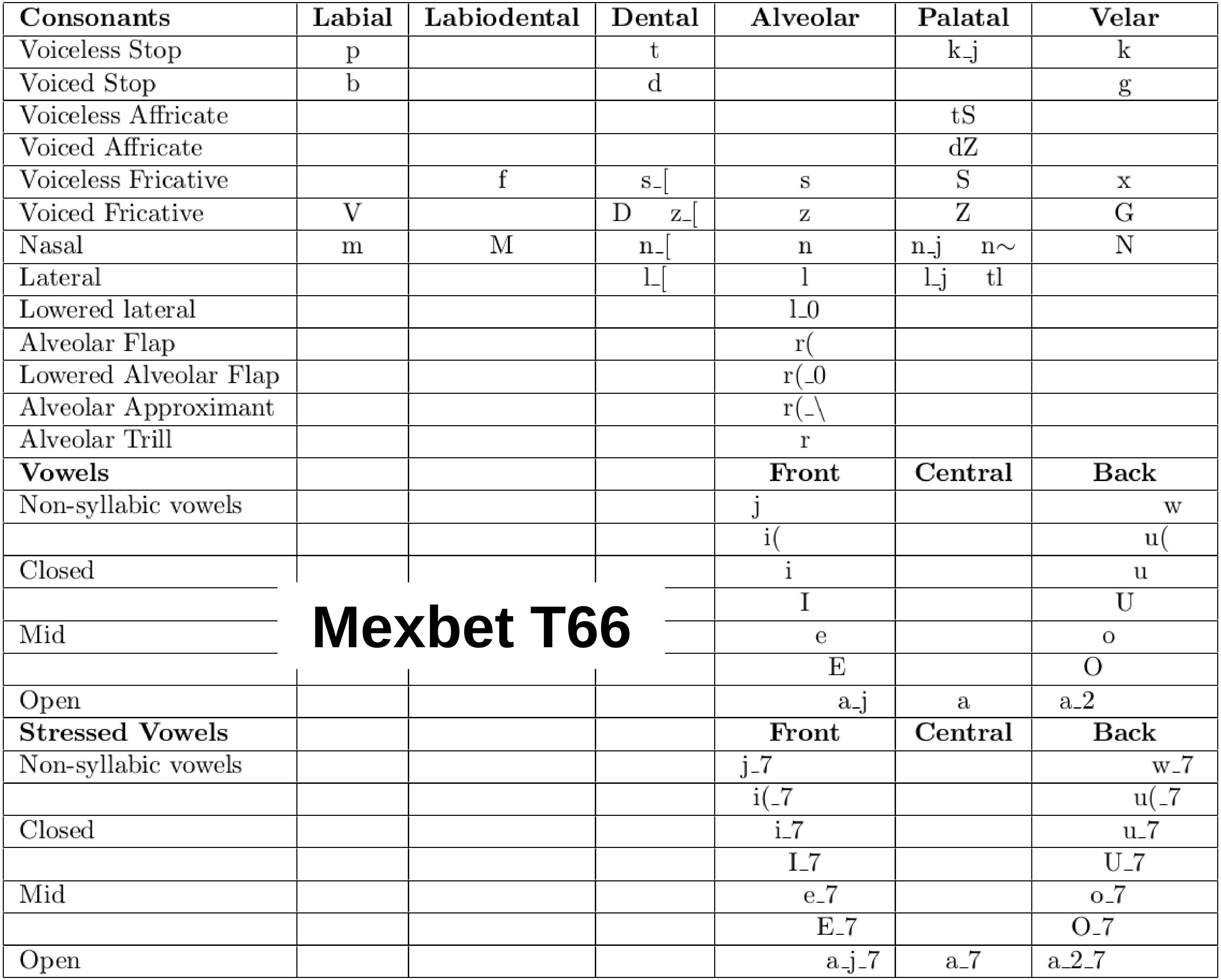
Mexbet T29



Consonants	Labial	Labiodental	Dental	Alveolar	Palatal	Velar
Voiceless Stop	p		t			k
Voiced Stop	b		d			g
Voiceless Affricate					tS	
Voiceless Fricative		f		s	S	x
Voiced Fricative					Z	
Nasal	m			n	n~	
Flap				r(		
Trill				r		
Lateral				l	tl	
Vowels				**Front**	**Central**	**Back**
Close				i		u
Mid				e		o
Open					a	
Stressed Vowels				**Front**	**Central**	**Back**
Close				i_7		u_7
Mid				e_7		o_7
Open					a_7	

Mexbet was created by Javier Cuétara in 2004

Mexbet T66



Consonants	Labial	Labiodental	Dental	Alveolar	Palatal	Velar
Voiceless Stop	p		t		k_j	k
Voiced Stop	b		d			g
Voiceless Affricate					tS	
Voiced Affricate					dZ	
Voiceless Fricative		f	s_[	s	S	x
Voiced Fricative	V		D z_[	z	Z	G
Nasal	m	M	n_[	n	n_j n~	N
Lateral			l_[	l	l_j tl	
Lowered lateral				l_0		
Alveolar Flap				r(		
Lowered Alveolar Flap				r(_0		
Alveolar Approximant				r(_\		
Alveolar Trill				r		
Vowels				**Front**	**Central**	**Back**
Non-syllabic vowels				j		w
				i(		u(
Closed				i		u
				I		U
Mid				e		o
				E		O
Open				a_j	a	a_2
Stressed Vowels				**Front**	**Central**	**Back**
Non-syllabic vowels				j_7		w_7
				i(_7		u(_7
Closed				i_7		u_7
				I_7		U_7
Mid				e_7		o_7
				E_7		O_7
Open				a_j_7	a_7	a_2_7

Problems with the Original CIEMPIESS

- Transcripts in a cryptic format (cAAsa, pEEro, gAAto,etc.)
- The train/test division was a bad idea.
- The timestamps and the orthographic transcriptions does not always match.
- It is not gender balanced
- It is not divided by speaker

The CIEMPIESS LIGHT Corpus

Characteristics

- Published at LDC in November 2017 (LDC2017S23).
- 18 hours long. No train/test division
- The transcriptions were revised.
- It doesn’t have timestamps.
- The transcriptions are orthographic only.
- Few female speakers were added but still gender unbalanced.
- The file names are very informative.

CMPL_F_01_03MAB_00007

This is SCLITE friendly!!!

New corpus to come…

The CIEMPIESS BALANCE Corpus

Characteristics

- It will be published soon at the LDC
- 18 hours not divided in train/test sets.
- Human transcriptions.
- Transcriptions in an orthographic level.
- In combination with the CIEMPIESS LIGHT one can have a perfectly gender-balanced corpus.
- Same file naming conventions as in the CIEMPIESS LIGHT

CMPB_M_01_01IVN_00001

The CIEMPIESS Experimentation Package

It includes:

- CIEMPIESS COMPLEMENTARY
- CIEMPIESS FEM
- CIEMPIESS TEST

It also includes:

- The fonetica3 library
- Documentation that explains how to create accurate transcripts in Mexican Spanish using the Mexbet phonetic alphabet.

The CIEMPIESS COMPLEMENTARY Corpus

Characteristics

- It is a 1 hour, gender-balanced corpus of people reading isolated Spanish words.
- It was designed to prevent any lack of phone instances when working with ASR systems.
- It is made by 10 women and 10 men reading words, digits and the alphabet.
- It includes two pronuncing diccionaries in Mexbet.
- Succesfully tested with Kaldi and PocketSphinx4
- Same file naming conventions as in the CIEMPIESS LIGHT.

CMPC_M_05_W_0001

The CIEMPIESS FEM Corpus

Characteristics

- 14 hours of female voices.
- Human transcriptions. Orthograpic level.
- Same file naming conventions as in the CIEMPIESS LIGHT.

CMPL_F_08_04ALX_00008

The CIEMPIESS TEST Corpus

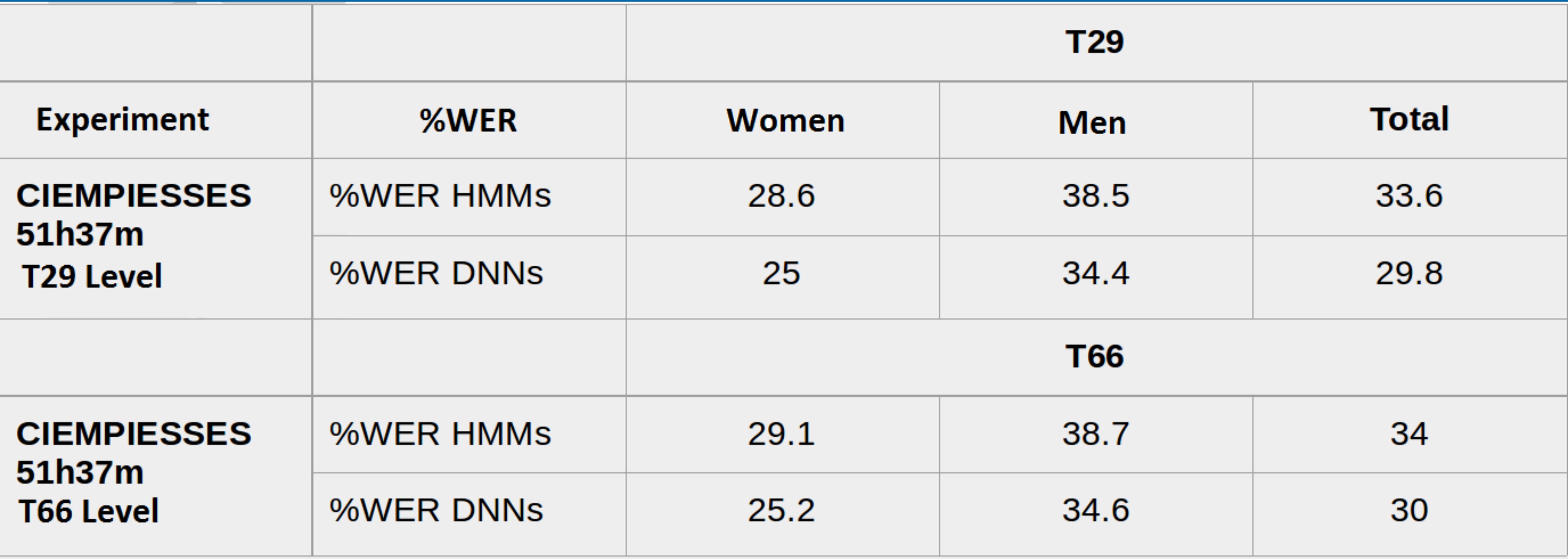
Characteristics

- It is an 8 hours gender balance corpus (10 males and 10 females).
- Human transcriptions. Orthograpic level..
- Same file naming conventions as in the CIEMPIESS LIGHT.

CMPT_F_01_0001

ASR Experiment with Kaldi

(Using Mexbet)



		T29		
Experiment	**%WER**	**Women**	**Men**	**Total**
CIEMPIESSES 51h37m T29 Level	%WER HMMs	28.6	38.5	33.6
	%WER DNNs	25	34.4	29.8
		T66		
CIEMPIESSES 51h37m T66 Level	%WER HMMs	29.1	38.7	34
	%WER DNNs	25.2	34.6	30

Thanks for your attention

CONTACT:

Carlos Daniel Hernández Mena

ca_hernandez@uxmcc2.iimas.unam.mx

www.ciempiess.org